Theoretical Computer Science 454 (2012) 5–22

Contents lists available at SciVerse ScienceDirect

Theoretical Computer Science

journal homepage: www.elsevier.com/locate/tcs

Rewriting rule chains modeling DNA rearrangement pathways

Angela Angeleska [a,*], Nataša Jonoska [b], Masahico Saito [b]

[a] Department of Mathematics, University of Tampa, United States

[b] Department of Mathematics and Statistics, University of South Florida, United States

ARTICLE INFO

Keywords:
Assembly pathways
DNA rearrangement
Rewriting rules
Ciliates

ABSTRACT

We introduce a model that describes rearrangement pathways of DNA recombination events. The recombination processes may happen in a succession, possibly with some recombination events performed simultaneously, but others in a prescribed order. These events are modeled by three rewriting rules applied on a set of formal linear and circular words. We define a partial order on these sets in such a way that two sets are related by this order when molecules represented by one are produced by recombination events from the other. We apply our model to experimental data obtained for DNA rearrangement of the actin I gene in *O. trifallax* ciliates, and we predict possible pathways of gene rearrangement compatible with the data.

1. Introduction

Gene and genome rearrangements have been observed in a variety of cells, including unicellular eukaryotes (ciliates [8,22,24]) and cancer cells [5,7]. The best studied example of specific DNA excision and rearrangement is the processing of immunoglobulin and T-cell receptor genes in mammalian cells [12,13]. Some of the most excessive gene rearrangements and DNA recombination events appear in some species of ciliates, a group of eukaryotic unicellular organisms. Certain species, such as *Oxytricha nova* (also called *Sterkiella nova*) and *Stylonychia lemnae*, undergo complex DNA rearrangement (see [8,20,21]), which makes them ideal model organisms for studying these processes. Ciliates possess two types of nuclei: macronuclei (MAC) and micronuclei (MIC). Micronuclear genes are interrupted by non-coding segments (internal eliminated sequences, IESs), which divide a gene into segments called macronuclear destined sequences (MDSs, for short). The MDSs are not just interrupted by IESs, but also appear in scrambled order and may be inverted. During conjugation, haploid macronuclei are disintegrated and micronuclei exchanged. Subsequently, one of the micronuclei develops into a macronucleus which may afterwards produce several copies of itself. In this process the micronuclear DNA is transformed into macronuclear DNA through massive DNA rearrangement involving MDS unscrambling, IESs removal, and MDS inversion. This DNA recombination is assumed to be guided by repetitive sequences appearing at the end of the $(i-1)$th and the beginning of the ith MDS, called pointers, and theoretical models were proposed in [11,15,16].

The micronuclear actin I gene of *Oxytricha trifallax* (also known as *Sterkiella histriomuscorum*) is schematically represented in Fig. 1. In the figure, dark segments represent MDS sequences, with the ith MDS labeled by M_i for $i = 1, \ldots, 10$, and white segments represent IESs. The MAC gene consists of the correctly unscrambled MDS sequences $(M_1 M_2 \cdots M_{10})$ with some IESs at the both ends, and some circular IES molecules that are eliminated.

Models proposing involvement of a new molecule guiding the recombination, called a template, are proposed in [3,23]. Nowacki et al. in [18] proved the involvement of templates in *O. trifallax* by experimentally directing DNA unscrambling with

* Corresponding author. Tel.: +1 8133176793.
E-mail addresses: aangeleska@ut.edu (A. Angeleska), jonoska@math.usf.edu (N. Jonoska), saito@math.usf.edu (M. Saito).

0304-3975/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2012.04.041

Fig. 1. Schematic representation of the actin I micronuclear gene in *O. trifallax*.

synthetic RNA templates. For a more detailed description of the biological phenomenon we refer to [20–22] and references therein.

The authors of [17] experimentally confirmed that the unscrambling of the MIC genes is a cascading process. Namely, it was observed that (1) some recombination events take place in a preferred order, (2) different recombination processes, called pathways, seem to happen in parallel, and (3) during the rearrangement the MDSs may be separated such that they appear in two or more distinct molecules (see also Section 5 for details). The fact that the MDSs may be separated during the rearrangement suggests that the process includes multiple molecules, which was not considered, for example, in [4,11]. However, there is no evidence that portions of the molecules from one copy of a partially assembled gene exchange with portions from another copy of a partially assembled gene; hence the process may not be completely intermolecular as proposed in [15,16]. These results motivate our mathematical model in which, at each instance of a macronuclear gene assembly, all resulting potential molecules are kept in one set and only the molecules within this set are allowed to interact further. The set of different molecules present at a certain instance of the recombination is called a *set of intermediates* that consists of linear and circular words. The DNA rearrangement is formalized by introducing three types of transformation (rewriting rules) on sets of intermediates: *deletion, insertion*, and *inversion*. These rewriting rules are very similar to the operations defined in [15,16].

We describe consecutive steps in the process of gene assembly by ordering sets of intermediates. Namely, a strict partial order is defined on the sets of intermediates, such that two sets of intermediates are related by this order if one set of intermediates can be obtained from the other by application of a composition of deletion, insertion, or inversion operations that increases the size of partially assembled segments. We define an assembly strategy to be a linearly ordered subset of sets of intermediates, in which the minimal element corresponds to the scrambled micronuclear gene and the maximal element corresponds to the assembled macronuclear gene. In this model, we show that, for any set of intermediates that models a MIC gene, there is an assembly strategy that will reduce it to a maximal set of intermediates that contains a MAC gene. Furthermore, we prove in Theorem 3.11 that the rewriting system is confluent. In other words, for a given minimal element, regardless of the pathway (assembly strategy), the maximal element (assembled product and all other resulting sequences) is unique. This result allows an algorithm given in Section 3, which outputs the unique maximal set of intermediates (representing a MAC gene with other possible molecules composed of IESs), obtained by rearrangement of an input minimal set of intermediates (MIC gene).

The rewriting system on sets of intermediates is applied on experimental data in Section 5. Based on the experimental data in [17], we postulate two different pathways for descrambling the actin I gene of *O. trifallax*, each modeled through an assembly strategy. We view each assembly strategy as a possible pathway of a gene rearrangement.

Furthermore, the mathematical model introduced here was a motivation for the experimental work presented in [2], where different rearrangement strategies were generated for the TEBPα gene in *O. trifallax*, and putative intermediate molecules containing multiple IESs were observed for the first time.

Detailed comparison of the model presented here with the intramolecular model in [11] and the intermolecular model in [15,16] can be found in [1] (see also the concluding remarks).

2. Definitions and notation

We use notation M_i to denote the ith MDS and I_j to denote the jth IES in a micronuclear sequence. Let $A_k = \{M_1, \ldots, M_k, I_0, \ldots, I_k\}$ and $\overline{A_k} = \{\overline{M_1}, \ldots, \overline{M_k}, \overline{I_0}, \ldots, \overline{I_k}\}$ be disjoint alphabets (sets of symbols) for some integer $k \geq 1$, and let $\mathfrak{A}_k = A_k \cup \overline{A_k}$. We also write $\mathcal{M} = \{M_1, \ldots, M_k, \overline{M_1}, \ldots, \overline{M_k}\}$ and $\mathcal{I} = \{I_0, \ldots, I_k, \overline{I_0}, \ldots, \overline{I_k}\}$. Define an involution $\theta : \mathfrak{A}_k \to \mathfrak{A}_k$, $\theta^2 = \mathrm{id}$, by $\theta(M_i) = \overline{M_i}$, $\theta(I_j) = \overline{I_j}$ for $1 \leq i, j \leq k$. We also use the notation $\overline{X}$ representing $\theta(X)$ for $X \in \mathfrak{A}_k$, thus $\overline{\overline{M_i}} = M_i$, for example. We extend θ to $\mathfrak{A}_k^*$, the set of all words over $\mathfrak{A}_k$, by the antimorphism property, so do not change that for a word $v = v_1 \cdots v_n$, where $v_i \in \mathfrak{A}_k$ and n is a positive integer, $\theta(v) = \theta(v_n) \cdots \theta(v_1)$, or $\overline{v} = \overline{v_n \cdots v_1}$. For a word v over $\mathfrak{A}_k$, the word $\theta(v) = \overline{v}$ is called the *reverse* of v.

2.1. Sets of intermediates

We consider two types of words over $\mathfrak{A}_k$: linear words that correspond to the regular definition of words over an alphabet, and circular words.

Definition 2.1. Two words v and w over $\mathfrak{A}_k$ are *equivalent* (as linear words) if $v = w$ or $v = \overline{w}$. Two words v and w over $\mathfrak{A}_k$ are *cyclically equivalent* if w is obtained from v by a sequence of cyclic permutations and applications of θ. A *circular word* over an alphabet $\mathfrak{A}_k$ is a cyclic equivalence class of words over $\mathfrak{A}_k$.

For a word w, we denote the circular word corresponding to w by $[w]$, but often abbreviate the notation and say "a circular word w" if no confusion arises. In this case, w is a representative word of the class $[w]$.

Example 2.2. For example, $I_0M_1I_1\overline{M_2}$ and $M_2\overline{I_1M_1I_0}$ are equivalent as linear words, and $\overline{M_1I_0}M_2\overline{I_1}$ is cyclically equivalent to both. Thus $[I_0M_1I_1\overline{M_2}] = [\overline{M_1I_0}M_2\overline{I_1}]$, and the circular word $I_0M_1I_1\overline{M_2}$ is the same as the circular word $\overline{M_1I_0}M_2\overline{I_1}$.

Definition 2.3. Let $k > 0, n \geq 0$ be integers. A *set of intermediates* over an alphabet $\mathfrak{A}_k$ is a set

$$W = \{ w_0, [w_1], [w_2], \ldots, [w_n] \},$$

where w_0 is a linear word, and $[w_1], [w_2], \ldots, [w_n]$ are n circular words over $\mathfrak{A}_k$ (empty if $n = 0$), such that, for each symbol $X \in \mathfrak{A}_k$, either X or $\overline{X}$ appears exactly once in W.

We say that two sets of intermediates are equal if they are equal as sets. The set of all sets of intermediates over alphabet $\mathfrak{A}_k$ is denoted by $\mathfrak{A}_k^{\sim}$.

A set of intermediates corresponds to a collection of DNA molecules present at a certain step during the rearrangement with the assumption that all of them are obtained as assembly (by)products of a single gene. Circular words represent potential circular DNA molecules, some of which have been observed experimentally [2].

Example 2.4. Let $k = 4$ and $w_0 = M_3M_4I_1\overline{M_2}I_3I_4$, $[w_1] = [I_0]$, $[w_2] = [M_1I_0]$, $[w_3] = [\overline{M_1}I_2]$ be words over $\mathfrak{A}_4$. Then $W = \{ w_0, [w_1], [w_3] \}$ is a set of intermediates, but $W' = \{ w_0, [w_2] \}$ is not, since I_2 does not appear in W'. Also, $W'' = \{ w_0, [w_2], [w_3] \}$ is not a set of intermediates, since M_1 appears in w_2 and $\overline{M_1}$ appears in w_3.

For a word s over $\mathfrak{A}_k$ and a set of intermediates $W \in \mathfrak{A}_k^{\sim}$, we denote $s \sqsubset W$ if there is an element $w \in W$ such that s is a substring of w. In other words, $s \sqsubset W$ if s is a substring of some word in W. Recall that $w \in W$ is, in fact, an equivalence class, so that $s \sqsubset W$ means that s is a substring of some representative of the class that w belongs to.

Definition 2.5. Each word over $\mathfrak{A}_k$ of the form $M_iM_{(i+1)}\cdots M_{(i+j)}$ or $\overline{M_{(i+j)}\cdots M_i}$, for some $i \in \{1, 2, \ldots, k\}$ and $j \in \{0, 1, \ldots, k-i\}$, is called an *assembled segment*.

An assembled segment $s \sqsubset W$ is *maximal* if, for every assembled segment $s' \sqsubset W$, if s is a substring of s', then $s' = s$. Note that a set of intermediates W can have multiple maximal assembled segments. We denote by S_W the (unordered) set of all maximal assembled segments in a set of intermediates W.

Example 2.6. For $W = \{ I_0M_1M_2I_4M_5I_5, [I_1M_3M_4I_3], [I_2] \}$, we have $S_W = \{ M_1M_2, M_3M_4, M_5 \}$.

Definition 2.7. For a word s, let $|s|$ denote the length (the number of symbols) of s. The *degree* of a set of intermediates W is defined as $\Sigma_{s \in S_W}(|s| - 1)$, and is denoted by $\mathrm{dg}(W)$.

The degree of a set of intermediates W counts the number of pairs $M_iM_{(i+1)}$ or $\overline{M_{(i+1)}M_i}$ that appear as subwords in the elements of W. The degree of a set of intermediates can also be seen as a number of MDS junctions within a partially assembled DNA molecule. The degree of the set of intermediates in Example 2.6 is 2.

Definition 2.8. A set of intermediates W is called *realizable* if $NN' \sqsubset W$ for $N, N' \in \mathcal{M}$ implies that $NN' = M_iM_{i+1}$ or $NN' = \overline{M_{i+1}M_i}$, where $i = 1, \ldots, k-1$.

The word *realizable* is motivated from *realizable words* discussed in [11]. Realizable sets of intermediates model MAC genes, MIC genes, or correctly partially rearranged molecules. Studying sets of intermediates that are not realizable has biological relevance as well. It is experimentally observed in [2,17], that some of the partially assembled molecules are not correctly assembled, i.e., they can be aberrant. The aberrant intermediates can be modeled by non-realizable sets of intermediates.

2.2. Rewriting rules

In this section, we define rewriting rules (operations) and investigate their properties. We define three types of transformation (rewriting rules) on sets of intermediates: deletion, insertion, and inversion. By the definition of circular words, any class representative of a circular word can be used.

Definition 2.9. Let $W = \{w_0, [w_1], \ldots, [w_n]\}$ be a set of intermediates over alphabet $\mathfrak{A}_k$. Let $u, v, w \in \mathfrak{A}_k^*$.

- Deletion (Fig. 2):
  Let $w_j = uwv$ and $w_j' = uv$. We say that the set of intermediates W' is obtained from W by *deletion* and write $W \overset{del}{\Rightarrow} W'$ if

$$W' = \begin{cases} (W \setminus \{[w_j]\}) \cup \{[w_j'], [w]\} & \text{if } j \neq 0, \\ (W \setminus \{w_0\}) \cup \{w_0', [w]\} & \text{if } j = 0. \end{cases}$$

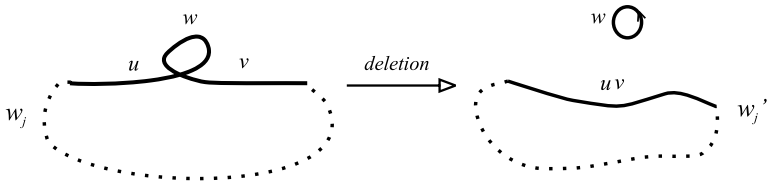

Fig. 2. Deletion.

- Insertion (Fig. 3):
  Let $w_j = uv$, where u or v may be empty, and $w'_j = uw_iv$ for $i > 0, i \neq j$. We say that the set of intermediates W' is obtained from W by *insertion* and write $W \overset{ins}{\Rightarrow} W'$ if

$$W' = \begin{cases} (W \setminus \{[w_j], [w_i]\}) \cup \{[w'_j]\} & \text{if } j \neq 0, \\ (W \setminus \{w_0, [w_i]\}) \cup \{[w'_0]\} & \text{if } j = 0. \end{cases}$$

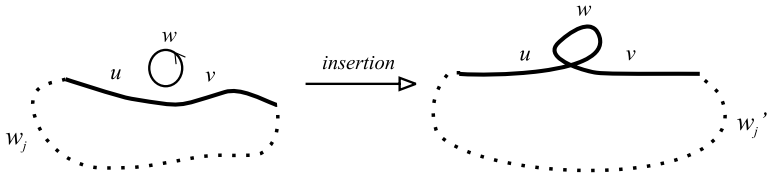

Fig. 3. Insertion.

- Inversion (Fig. 4):
  Let $w_j = uwv$ and $w'_j = u\overline{w}v$. We say that the set of intermediates W' is obtained from W by *inversion* and write $W \overset{inv}{\Rightarrow} W'$ if

$$W' = \begin{cases} (W \setminus \{[w_j]\}) \cup \{[w'_j]\} & \text{if } j \neq 0, \\ (W \setminus \{w_0\}) \cup \{w'_0\} & \text{if } j = 0. \end{cases}$$

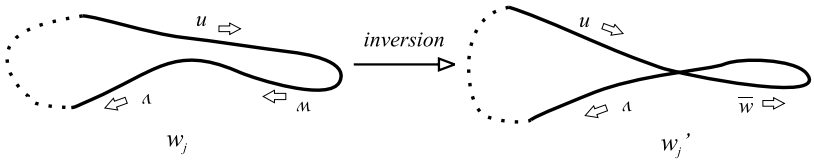

Fig. 4. Inversion.

If W' is obtained from W by any of the three operations (rewriting rules) deletion, insertion, or inversion, we write $W \overset{r}{\Rightarrow} W'$. Directly from the definition we have that $W \overset{ins}{\Rightarrow} W' \Leftrightarrow W' \overset{del}{\Rightarrow} W$ and that $W \overset{inv}{\Rightarrow} W' \overset{inv}{\Rightarrow} W$, when the operations are performed to the appropriate words.

The deletion and insertion operations defined on assembly words are similar to the operations $op1$ and $op1^R$, respectively, from the intermolecular model introduced by Kari and Landweber in [15,16], while inversion operation is the same as the hairpin recombination from the intramolecular model introduced by Rozenberg et al. in [10,11,19]. Even though parts of operations in [10,11,15,16,19] overlap with our model, our approach differs from both a theoretical and a biological perspective. The differences can be briefly summarized as follows: we do not apply the rewriting rules to words coming from different sets of intermediates (in contrast to the intermolecular model approach [15]), but we allow rewriting rules applied to different words within the same set of intermediates (in contrast to the intramolecular model approach [11]). Further similarities and differences are discussed in [1].

The following lemma shows that the set of sets of intermediates is closed under deletion, insertion, and inversion.

Lemma 2.10. *The set W' obtained from a set of intermediates W by a deletion, insertion, or inversion, is a set of intermediates.*

Proof. Let $W \overset{r}{\Rightarrow} W'$, where $r \in \{del, ins, inv\}$. Note that the number of linear words in W' remains unchanged. Further, if r is insertion or deletion, then a subword of a word in W is relocated as a subword of possibly another word in W'; hence the number of symbols and the number of their appearances in W and those in W' are equal. If r is an inversion performed on $w \sqsubset W$, then $\overline{w} \sqsubset W'$, and all other symbols and words remain unchanged. Hence, for any r, W is a set of intermediates if and only if W' is a set of intermediates. □

3. Assembly strategies

In this section, we define a partial order on the sets of intermediates and consider the chains that contain the minimal and the maximal elements. These chains correspond to possible assembly pathways in the gene assembly.

3.1. Partial order

In this section, we present definitions and basic properties.

Definition 3.1. Given an alphabet $\mathfrak{A}_k$, we define a binary relation $\Rightarrow^*$ on the family of sets of intermediates $\mathfrak{A}_k^{\sim}$ over $\mathfrak{A}_k$ as follows: $W \Rightarrow^* W'$ if there is a sequence $W = W_0, W_1, \ldots, W_h = W'$ of sets of intermediates such that, for each i $(i = 1, \ldots, h)$, W_i is obtained from W_{i-1} by a single operation of deletion, insertion, or inversion, and $\mathrm{dg}(W_{i-1}) < \mathrm{dg}(W_i)$.

Since $<$ is a strict partial order on non-negative integers, it follows directly from the definition that $\Rightarrow^*$ is a strict partial order.

Lemma 3.2. *For a given alphabet $\mathfrak{A}_k$, the sets of intermediates of degree 0 are minimal elements and the sets of intermediates of degree $k-1$ are maximal elements of $(\mathfrak{A}_k^{\sim}, \Rightarrow^*)$.*

Proof. Since $W \Rightarrow^* W'$ implies that $\mathrm{dg}(W) < \mathrm{dg}(W')$, sets of intermediates that have degree 0 are minimal and sets of intermediates that have degree $k-1$ are maximal. □

Example 3.3. Let $W = \{I_0M_3I_1M_4I_2\overline{M_2}I_3M_1I_4\}$ be a set of intermediates over $\mathfrak{A}_4$. Then we have

$$
\begin{aligned}
(i)\quad W &\overset{del\ I_1}{\Rightarrow} \{I_0M_3M_4I_2\overline{M_2}I_3M_1I_4, [I_1]\} \overset{inv\ I_3M_1}{\Rightarrow} \{I_0M_3M_4I_2\overline{M_2M_1I_3}I_4, [I_1]\}\\
&\overset{inv\ M_3M_4I_2}{\Rightarrow} \{I_0\overline{I_2M_4M_3M_2M_1I_3}I_4, [I_1]\} = W',\\
(ii)\quad W &\overset{inv\ I_3M_1}{\Rightarrow} \{I_0M_3I_1M_4I_2\overline{M_2M_1I_3}I_4\} \overset{del\ I_1}{\Rightarrow} \{I_0M_3M_4I_2\overline{M_2M_1I_3}I_4, [I_1]\}.
\end{aligned}
$$

The notation above $\Rightarrow$ indicates the operation types and the subwords on which they were performed. Note that W is a minimal element and that W' is a maximal element.

Definition 3.4. An *assembly strategy* is a linearly ordered subset (chain) of $(\mathfrak{A}_k^{\sim}, \Rightarrow^*)$ that contains a maximal and a minimal element of $(\mathfrak{A}_k^{\sim}, \Rightarrow^*)$.

In Example 3.3, chain (i) is an assembly strategy, but chain (ii) is not.

The condition $\mathrm{dg}(W_{i-1}) < \mathrm{dg}(W_i)$ in the definition of $\Rightarrow^*$ is based on the assumption that in each step of the rearrangement MDSs are joined together, and once joined they cannot be separated. In other words, we suppose that the recombination is irreversible. If we consider the DNA rearrangement in ciliates to be guided solely by pointers, then we would have to assume reversibility, as, after recombination, the pair of pointers is still present in the set of intermediates. Our motivation for strict ordering considers possible templates involved in the process [3,18,23]. Although the pairs of recombined pointers are present, their left and right contexts are different than before the recombination, and therefore the template-guided recombinations may not be repeated.

The minimal elements in an assembly strategy correspond to MIC gene sequences while the maximal elements correspond to the sets of molecules including the correctly assembled MAC gene together with molecules composed of excised IESs. In what follows, we show that every assembly strategy with a given minimal element contains a unique maximal element (up to equivalence).

Proposition 3.5. *For every set of intermediates W with $\mathrm{dg}(W) < k-1$ over $\mathfrak{A}_k$, there is a rewriting rule r (deletion, insertion, or inversion) applicable to W such that $\mathrm{dg}(W') = \mathrm{dg}(W) + 1$, where $W \overset{r}{\Rightarrow} W'$.*

Proof. Let $W = \{w_0, [w_1], \ldots, [w_n]\}$ be a set of intermediates over alphabet $\mathfrak{A}_k$ such that $\mathrm{dg}(W) = d < k-1$. There is a symbol M_i in $\mathfrak{A}_k$ such that $M_iM_{i+1} \not\sqsubset W$ and $\overline{M_{i+1}M_i} \not\sqsubset W$ (if there is no such symbol then $M_1M_2 \ldots M_k$ or $\overline{M_k \cdots M_1} \sqsubset W$ and $\mathrm{dg}(W) = k-1$). There are two possibilities: either both M_i and M_{i+1} belong to a single word, or they belong to different words of W.

First, assume that both M_i and M_{i+1} belong to some $w_j \in W$. There are four cases to consider up to equivalence:

$$
\text{(a) } w_j = v_1M_iv_2M_{i+1}v_3, \qquad \text{(b) } w_j = v_1M_{i+1}v_2M_iv_3, \qquad \text{(c) } w_j = v_1M_iv_2\overline{M_{i+1}}v_3, \qquad \text{(d) } w_j = v_1\overline{M_i}v_2M_{i+1}v_3,
$$

where v_i, $i = 1, 2, 3$, are words over $\mathfrak{A}_k$, and v_2 is not the empty word in case (a). For a circular word w_j $(j \neq 0)$, cases (a) and (b), and cases (c) and (d), are equivalent.

(a) Let $w' = v_1M_iM_{i+1}v_3$. Then $W \overset{inv}{\Rightarrow} W' = (W \setminus \{[w_j]\}) \cup \{[w'], [v_2]\}$ is obtained from W by deletion and $\mathrm{dg}(W') = d+1$ if $j \neq 0$. Similarly for $j = 0$.

(b) Let $w' = v_1v_3$. Then $W \overset{del}{\Rightarrow} W' = (W \setminus \{[w_j]\}) \cup \{[w'], [M_{i+1}v_2M_i]\}$ is obtained from W by deletion with $\mathrm{dg}(W') = d+1$, since $[M_{i+1}v_2M_i]$ equals $[M_1M_{i+1}v_2]$.

(c) In this case, $W' = \{w_0, [w_1], \ldots, [w_{j-1}], [v_1 M_i M_{i+1} \overline{v_2} v_3], [w_{j+1}], \ldots, [w_n]\}$ is a set of intermediates with degree $d+1$, obtained from W by inversion of $v_2 \overline{M_{i+1}}$. Similarly, in case (d), there is a set of intermediates W' with degree $d+1$ obtained from W by an inversion of $\overline{M_i} v_2$.

Next, assume that M_i and M_{i+1} belong to separate words in W. If M_i appears in a circular word, we can always choose a representative of the word that starts with M_i. There are three different cases up to equivalence:

(a) $w_0 = v_1 M_i v_2$ and $w_j = [M_{i+1} v_3]$,
(b) $w_0 = v_1 M_{i+1} v_2$ and $w_j = [M_i v_3]$,
(c) $w_s = [M_i v_1]$ and $w_j = [M_{i+1} v_2]$,

where v_i $(i = 1, 2, 3)$ are words over $\mathfrak{A}_k$ and w_j, w_s are circular words of W, $0 < j, s \leq n$. In all three cases, w_j can be inserted in W such that $M_i M_{i+1} \sqsubset W'$, where $W \overset{ins}{\Rightarrow} W'$. □

As a direct consequence of Proposition 3.5, we have the following theorem.

Theorem 3.6. *For every minimal set of intermediates W over $\mathfrak{A}_k$, there is an assembly strategy that contains W.*

Proposition 3.7 justifies the necessity of all three rewriting operations (insertion, deletion, and inversion) by proving that, in general, none of them can be a combination of the other two.

Proposition 3.7. *If one of the rules insertion, deletion, inversion is excluded, then there is a minimal set of intermediates W for which there is no assembly strategy containing W.*

Proof. Consider the following singleton sets of intermediates of degree 0:

$W_1 = \{I_1 M_1 I_2 M_2 I_3\}$,
$W_2 = \{I_1 M_1 I_2 \overline{M_2} I_3\}$,
$W_3 = \{I_1 M_1 I_2 M_3 I_3 M_2 I_4\}$.

Observe that, in an assembly strategy, every step of $\Rightarrow^*$ requires an increase in the degree. So any single operation that is performed in an assembly strategy containing W_1 or W_2 must produce $M_1 M_2$ as a subword. For W_1, this is possible only if I_2 is deleted, and for W_2 only if $I_2 \overline{M_2}$ is inverted.

A single operation applied to W_3 requires that at least one of $M_1 M_2$, $M_2 M_3$ appears as a subword in the resulting set of intermediates. There are two possibilities: $W_3 \overset{del}{\Rightarrow} W_3' = \{I_1 M_1 M_2 I_4, [I_2 M_3 I_3]\}$ or $W_3 \overset{del}{\Rightarrow} W_3'' = \{I_1 M_1 I_2 I_4, [M_3 I_3 M_2]\}$. In each case, only insertion of the circular word completes the assembly strategies. □

3.2. Confluence

In this section, we show that the rewriting model defined in Section 2.2 is confluent (Theorem 3.11). In terms of DNA assembly, it means that, regardless of the assembly pathway, the resulting set of intermediates (including the correctly assembled MAC gene) is always the same.

Definition 3.8. Let W be a set of intermediates over alphabet $\mathfrak{A}_k$. Then, for every $N \in \mathfrak{A}_k$, there are $X, Y \in \mathfrak{A}_k \cup \{\epsilon$ (the empty word)$\}$ such that $XNY \sqsubset W$. We say that X is the *left context* of N in W, and that Y is the *right context* of N in W.

We remark that the notions of left and right contexts are defined up to equivalence, so that, for a symbol N appearing in a word w, the left and the right contexts of N are the same in every representative of the equivalence class of w. In particular, if $X\overline{N}Y \sqsubset W$, then $\overline{Y}$ is the left context of N and $\overline{X}$ is the right context of N. Also, if $[X v Y]$ is a circular word such that $X, Y \in \mathfrak{A}_k$, then Y is a left context of X and Y is a right context of X.

Remark 3.9. If W is a minimal realizable set of intermediates, then the right and the left contexts of every $M_i \in \mathfrak{A}_k$ belong to $\{I_0, \ldots I_k, \overline{I_0}, \ldots, \overline{I_k}\}$.

We use Lemma 3.10 to prove Theorem 3.11.

Lemma 3.10. *Let W' and W'' be two sets of intermediates over alphabet $\mathfrak{A}_k$. If, for every symbol N, the right and left contexts of N in W' equal the right and the left contexts of N in W'', respectively, then W' is equal to W''.*

Proof. By definition, W' and W'' are composed of one linear word and multiple or no circular words. Let w_0' and w_0'' be the linear words in W' and W'', respectively. First, we show that w_0' is equal to w_0'' or its reverse. Let S_1 be the first symbol of w_0'. The left context of S_1 in W' is ϵ, and S_1 is the only symbol that has left context ϵ in W'. On the other hand, S_1 or $\overline{S_1}$ appears in W'', and S_1 has the same left context in W'' as in W'. That context is ϵ, which is possible only if

(1) S_1 is the first symbol of w_0'', or
(2) $\overline{S_1}$ is the last symbol in w_0''.

In case (1), the first symbols w_0' and w_0'' are the same. Inductively, assume that w_0' and w_0'' agree on the first n symbols, $S_1, S_2, \ldots, S_n$. The right context of S_n in w_0' is S_{n+1} if and only if the right context of S_n in w_0'' is S_{n+1}, implying that w_0' and w_0'' agree on the $(n+1)$th symbol. Hence $w_0' = w_0''$. Case (2) follows similarly.

Let w_j' $(0 < j \leq k-1)$ be any circular word in W', and let S be a symbol that belongs to w_j'. Then S or $\overline{S}$ belongs to a circular word w_j'' in W''. Consider the representatives of w_j' and w_j'' that start with S and pursue induction as in case (1) above. □

The following theorem comes as no surprise, as confluences have been proven for other rewriting models (e.g., [9–11]).

Theorem 3.11. *Let W be a minimal realizable set of intermediates over alphabet $\mathfrak{A}_k$. There is a unique maximal set of intermediates W' such that every assembly strategy that contains W as a minimal set of intermediates contains W'.*

Proof. Let W be a set of intermediates and $S : W = W_0, W_1, W_2, \ldots, W_h = W'$ be an assembly strategy, where W' is a maximal set of intermediates. Each W_s for $s \in \{1, 2, \ldots, h\}$ is obtained from W_{s-1} by application of one or more insertion, inversion, or deletion operations. Therefore, we can think of W' as a set of intermediates obtained from W by a sequence $r_1, r_2, \ldots, r_{k-1}$ of inversion, deletion, or insertion rules, such that every application of a rule increases the degree by one. Let $W \overset{r_1}{\Rightarrow} R_1 \overset{r_2}{\Rightarrow} \cdots \overset{r_{k-1}}{\Rightarrow} R_{k-1} = W'$. Note that the left and the right context of every M_i in the set of intermediates changes between R_{j-1} and R_j if and only if r_j joins M_i with either M_{i-1} or M_{i+1} to form $M_{i-1}M_i$ or M_iM_{i+1}, respectively. Suppose that r_j is a rule such that $M_iM_{i+1} \not\sqsubset R_{j-1}$ and $M_iM_{i+1} \sqsubset R_j$ or $\overline{M_{i+1}M_i} \not\sqsubset R_{j-1}$ and $\overline{M_{i+1}M_i} \sqsubset R_j$. Let Y be the right context of M_i and X be the left context of M_{i+1}. Then, regardless of the type of rule r_j, we show that the following holds:

(1) $Y \neq X$, and either $XY \sqsubset R_j$ or $\overline{YX} \sqsubset R_j$, or
(2) $Y = X$ and $[Y] \in R_j$.

This is shown by checking each case in the proof of Proposition 3.5 as follows.

First, assume that both M_i and M_{i+1} belong to some $w_j \in R_{j-1}$. Suppose that $j = 0$ (w_0 is a linear word). Then there are four cases, as in the proof of Proposition 3.5:

(a) $w_j = v_1M_iYv_2XM_{i+1}v_3$,
(b) $w_j = v_1XM_{i+1}v_2M_iYv_3$,
(c) $w_j = v_1M_iYv_2\overline{M_{i+1}X}v_3$,
(d) $w_j = v_1\overline{YM_i}v_2XM_{i+1}v_3$,

where v_i $(i = 1, 2, 3)$ represent some words.
(a) Suppose that $X \neq Y$. Since $M_iM_{i+1} \sqsubset R_j$, r_j must be a deletion of $[Yv_2'X]$; hence condition (1) holds. If $X = Y$, then r_j is a deletion of $[Y]$.
(b) In this case, $X \neq Y$, and $w_{n+1} = [M_{i+1}v_2M_i]$ is deleted as a circular word to create a sequence $M_iM_{i+1} \sqsubset R_j$, so that $w_0' = v_1XYv_3$, and condition (1) holds.
(c) In this case, $X \neq Y$, and $R_j = \{w_0 = v_1M_iM_{i+1}\overline{v_2Y_iX_{i+1}}v_3, [w_1], \ldots, [w_n]\}$, where r_j is an inversion of $Yv_2\overline{M_{i+1}}$ and (1) holds. Case (d) is similar, where r_j inverts $\overline{M_i}v_2X$. The cases when $j \neq 0$ are similar.

The cases when M_i and M_{i+1} belong to distinct words are similarly checked, and we obtain conditions (1) and (2). Since the degree must increase by every operation in an assembly strategy, any subsequent operation affects contexts of a pair M_j, M_{j+1}, different than M_i, M_{i+1}, so that XY, $\overline{YX}$, or $[Y]$ remain intact. Therefore, XY or $\overline{YX}$ in case (1), and $[Y]$ in case (2), remain subwords of R_s for every $s = j, j+1, \ldots, k-1$ in addition to R_j. We conclude that, for a given minimal set of intermediates, every assembly strategy contains a maximal set of intermediates W' with the following properties.
(*) $M_1M_2 \cdots M_k \sqsubset W'$ or $\overline{M_kM_{k-1}} \cdots \overline{M_1} \sqsubset W'$, and either $XY \sqsubset W'$ or $\overline{YX} \sqsubset W'$ in case (1) or $[Y] \in W'$ in case (2) above is satisfied for every right context X of M_i and left context Y of M_{i+1}.

Let S' and S'' be two assembly strategies with a minimal set of intermediates W and with maximal set of intermediates W' and W'', respectively, satisfying the property (*). We show that every symbol in W' and W'' has the same left and right contexts, and therefore $W' = W''$ by Lemma 3.10.

Consider a symbol $M_i \in W', W''$. Then ,in both W' and W'', the right context of M_i must be M_{i+1} if $i \leq k-1$. By setting $j = 1$ and $j = k-1$ in (a)–(d) above, the left context of M_1 and the right context of M_k remain the same in both W' and W'.

Consider a symbol $I_j \in W', W''$ for some $j \in \{1, 2, \ldots k-1\}$. Then there are some i and $i' \in \{1, 2, \ldots k-1\}$ such that one of the following holds:

(i) $M_iI_jM_{i'} \sqsubset W$, or
(ii) $\overline{M_i}I_jM_{i'} \sqsubset W$, or
(iii) $M_iI_j\overline{M_{i'}} \sqsubset W$.

In case (i), I_j is the right context of M_i and the left context of $M_{i'}$. By (*), if $XI_jY \sqsubset W'$ for symbols X, Y (maybe empty), then X is the left context of M_{i+1} and Y is the right context of $M_{i'-1}$. The same argument applies to W'', so the left and right contexts of I_j agree for W' and W''. Similar arguments show cases (ii) and (iii). Hence, by Lemma 3.10, the maximal set of intermediates must be unique regardless of the assembly strategy. □

Next, we give an algorithm, called the algorithm for gene assembly, which outputs a maximal set of intermediates as a result of applying an arbitrary assembly strategy to an input minimal set of intermediates. This is possible, since Theorem 3.6 guarantees the existence of a maximal set of intermediates for any input minimal set of intermediates. Moreover, Theorem 3.11 proves that, for a given minimal set of intermediates, the maximal set of intermediates is unique up to equivalence. Since the (assembled) maximal set of intermediates is not strategy specific, the algorithm uses properties (1) and (2) in the proof of Theorem 3.11, which are invariant for any assembly strategy. In particular, the left and the right contexts of each M_i symbol in a set of intermediates is unchanged until a rewriting rule incorporating M_i is applied.

3.3. Algorithm for gene assembly

The algorithm is based on the observations discussed in the proof of Theorem 3.11. Note that a minimal set of intermediates consists of a single linear word over alphabet $\mathfrak{A}_k$. Let the input be a minimal set of intermediates $W = \{w_0\}$. We assume without loss of generality that w_0 starts from I_0 and ends with I_k. The algorithm constructs the elements s or $[s]$ of the unique maximal set of intermediates W', using property (1) in the proof of Theorem 3.11 that the left context of M_ℓ and the right context of $M_{\ell-1}$ appear consecutively in a maximal set of intermediates W'. Note that $M_1 \cdots M_k \sqsubset W'$ or $\overline{M_k \cdots M_1} \sqsubset W'$, so the main concern is the placement of symbols from $\mathfrak{I}$ within the linear and the circular words of W'.

Let J be the last letter of s, a partially constructed element of W'. Locate the right context M_ℓ of J. Then the algorithm locates the right context J' of $M_{\ell-1}$, and concatenates it to s to form sJ'. If J' appears in s, then the circular word $[s]$ is added to a partially constructed W'. If $J' = I_k$, the last symbol of w_0, then a linear word s is added. If M_1 or $\overline{M_k}$ is the right context of J, then $sJM_1M_2 \cdots M_kJ'$ or $s\overline{M_k \cdots M_1}J'$, respectively, replaces s in the algorithm. The algorithm halts when every symbol from $\mathfrak{I} = \{I_0, \ldots, I_k, \overline{I_0}, \ldots, \overline{I_k}\}$ is exhausted, and outputs W'.

The notation $s \leftarrow s'$ means that a word or a set s is replaced by s'. We use the convention that the symbol J is a symbol from $\mathfrak{I}$ and that N is a symbol from $\mathfrak{M}$.

Input: $w = I_0N_1I_1 \cdots N_kI_k$, where $N_i \in \mathfrak{M}$ and $I_j \in \mathfrak{I}$.

1. Set $J = I_0$, $s = J$, $U_I = \{I_0, \overline{I_0}\}$, and $W' = \emptyset$.
2. Find the right context N of J, and let ℓ be the subscript of N, i.e., $N = M_\ell$ or $\overline{M_\ell}$.
3. (a) If $N \neq M_1$ and $N \neq \overline{M_k}$, then find the right context $J' \in \mathfrak{I}$ of $M_{\ell-1}$, and set $J \leftarrow J'$. (Note that, if $\overline{M_{\ell-1}} \sqsubset W$ and I_j is the left context of $\overline{M_{\ell-1}}$, then $\overline{I_j}$ is the right context of $M_{\ell-1}$.)
   Set $s \leftarrow sJ$ if $J \notin U_I$. Otherwise, $W' \leftarrow W' \cup \{[s]\}$, and go to 6.
   (b) If $N = M_1$, then find the right context J of M_k. Set $s \leftarrow sM_1M_2 \cdots M_kJ$.
   (c) If $N = \overline{M_k}$, then find the right context J of $\overline{M_1}$. Set $s \leftarrow s\overline{M_k \cdots M_1}J$.
4. Set $U_I \leftarrow U_I \cup \{I_j, \overline{I_j}\}$, where $J = I_j$ or $\overline{I_j}$ for some j.
5. If $J = I_k$, then set $W' \leftarrow W' \cup \{s\}$, and go to 6. Otherwise, go to 2.
6. If $U_I \neq \mathfrak{I}$, then choose $I_m \in \{I_0, \ldots, I_k\} \setminus U_I$, and find the right context N of I_m and let q be the subscript of N, i.e., $N = M_q$ or $\overline{M_q}$. Otherwise, go to 9.
7. Set $s = I_m$ and $U_I \leftarrow U_I \cup \{I_m, \overline{I_m}\}$.
8. If $M_{q-1}I_mM_q \sqsubset W$ or $\overline{M_{q+1}I_mM_q} \sqsubset W$, then set $W' \leftarrow W' \cup \{[s]\}$ and go to 6. Otherwise, go to 3.
9. Output W'.

The algorithm above outputs the unique maximal set of intermediates, which is a result of the application of any assembly strategy to W. The algorithm provides the number of elements in W', i.e., $|W'|$, their structure, and the number of circular elements in W'.

Example 3.12. Consider the actin I gene of *O. trifallax* (see Fig. 1) [17] that corresponds to the minimal set of intermediates $W = \{I_0M_3I_1M_4I_2M_6I_3M_5I_4M_7I_5M_9I_6M_{10}I_7\overline{M_2}I_8M_1I_9M_8I_{10}\}$. The algorithm applied to W goes through the following steps.

Step 1 starts with $s = I_0$. Then, in step 2, we find M_3 as the right context of I_0, and $\ell = 3$. In step 3 (a), we find the right context of M_2, which is $\overline{I_7}$, since I_7 is the left context of $\overline{M_2}$. After execution of step 3, we obtain the string $s = I_0\overline{I_7}$, and, by step 4, $U_I = \{I_0, I_7, \overline{I_0}, \overline{I_7}\}$. In step 5, we return to step 2, since $J = I_7 \neq I_{10}$. Next, in step 2, the right context of $\overline{I_7}$ is $\overline{M_{10}}$, and $\overline{I_8}$ is the right context of $\overline{M_1}$. Hence, by repeatedly applying step 3 of the algorithm, we extend the string to $s = I_0\overline{I_7}(\overline{M_{10} \cdots M_1})\overline{I_8}$, and $U_I = \{I_0, I_7, I_8, \overline{I_0}, \overline{I_7}, \overline{I_8}\}$. After three more steps, the string $I_0\overline{I_7}(\overline{M_{10} \cdots M_1})\overline{I_8}I_9I_5I_{10}$ is obtained, and the process stops, since there is no M' such that $I_{10}M' \sqsubset W$. The string $s = I_0\overline{I_7}(\overline{M_{10} \cdots M_1})\overline{I_8}I_9I_5I_{10}$ is added to W'. At this point, $U_I = \{I_0, I_5, I_7, I_8, I_9, I_{10}, \overline{I_0}, \overline{I_5}, \overline{I_7}, \overline{I_8}, \overline{I_9}, \overline{I_{10}}\} \neq \mathfrak{I}$. If we choose $I_1 \notin U_I$, then $s = I_1$, and $U_I = \{I_0, I_1, I_5, I_7, I_8, I_9, I_{10}, \overline{I_0}, \overline{I_1}, \overline{I_5}, \overline{I_7}, \overline{I_8}, \overline{I_9}, \overline{I_{10}}\}$. Since I_1 is the right context of M_3 and the left context of M_4, $[I_1]$ is added to W'. Similarly, $[I_6]$ is added to W' and $U_I = \{I_0, I_1, I_5, I_6, I_7, I_8, I_9, I_{10}, \overline{I_0}, \overline{I_1}, \overline{I_5}, \overline{I_6}, \overline{I_7}, \overline{I_8}, \overline{I_9}, \overline{I_{10}}\}$.

If we choose $I_2 \notin U_I$, then we start building a new string from $s = I_2$. We obtain $I_2M_6 \sqsubset W$ and $M_5I_4 \sqsubset W$, and, therefore, $s = I_2I_4$. Then $I_4M_7 \sqsubset W$ and $M_6I_3 \sqsubset W$, and hence the string $s = I_2I_4I_3$ is constructed. We have $U_I = \mathfrak{I}$. The symbol I_3 is the right context of M_6; we are looking for the right context of M_5, which is I_4. Since I_4 is already in U_I, we add the circular string $[I_2I_4I_3]$ to W'. The output is $W' = \{I_0\overline{I_7}(\overline{M_{10} \cdots M_1})\overline{I_8}I_9I_5I_{10}, [I_2I_4I_3], [I_1], [I_6]\}$.

In the process of gene rearrangement, a molecule containing the correctly assembled actin I MAC gene with several molecules composed only of IESs is obtained. According to the algorithm, three potentially circular molecules are excised,

one composed only of IES_1, the second composed of IES_6, and the third composed of IES_2, IES_3 and IES_4. The left (right) context of the macronuclear actin I gene is expected to be $IES_0\overline{IES_7}$ ($\overline{IES_8}IES_9IES_5IES_{10}$), respectively.

4. Assembly graphs and sets of intermediates

In this section, we study how sets of intermediates and assembly strategies are related to assembly graphs and smoothing strategies introduced in [4]. We show that smoothing strategies correspond to some types of assembly strategies. The two models have different advantages; assembly graphs are helpful in visualizing the rearrangement process and quite intuitive, while sets of intermediates are more convenient for algorithmic descriptions of rearrangements. For example, Algorithm 3.3 is expressed in words and is ready to implement. To present a list of assembly strategies, in comparison, by drawing assembly graphs of intermediate steps would require "translations" into words to implement. Another motivation for the model on sets of intermediates is the necessity to represent intermediates which are not correctly assembled. Namely, a number of incorrectly processed molecules, which can also offer useful information on the rearrangement pathways, have been experimentally observed (see [2,17]). It is not possible to represent the aberrant molecules with a single assembly graph defined in this section, since a smoothing of a vertex can only produce correct joining of MDSs. On the other hand, we can arbitrarily apply the insertion, deletion, and inversion operations to a set of intermediates to explain the pathway to incorrectly assembled intermediates. As mentioned earlier, there are other string-based models for DNA rearrangement in ciliates. The sets of intermediates expressed in words (unlike the assembly graphs) are also suitable for comparisons with existing models such as those in [15,16,10,11].

4.1. Assembly graphs, transversals, and polygonal paths

Let $\Gamma = (V, E)$ be a finite graph with a set of vertices V and a set of edges E. We allow parallel edges and loops. Denote by $E(v)$ the set of edges that are incident to a vertex $v \in V$. Every loop at vertex v is counted as two different edges incident to v. The cardinality of $E(v)$ (counting loops twice) is called the *valency of* v.

Let $v \in V$ and $E(v) = \{e_1, \ldots, e_k\}$. For each v and an order $(e_{i_1}, \ldots, e_{i_k})$, we denote the corresponding circular string by $[e_{i_1} \cdots e_{i_k}]$. A *rigid vertex* is a pair $(v, [e_{i_1} \cdots e_{i_k}])$.

For a 4-valent rigid vertex $(v, [e_1e_2e_3e_4])$, we say that e_2 and e_4 are *neighbors with respect to* v to e_1 (and e_3) and, vice versa, e_1 and e_3 are neighbors to e_2 and e_4. Note that a loop is regarded as two different edges, which are neighbors of each other.

An *assembly graph* is a finite connected graph whose vertices are rigid vertices of valency 1 or 4. A vertex of valency 1 is called an *endpoint*. Note that the definition of assembly graph implies that the number of endpoints is always even. Two assembly graphs are *isomorphic* if they are isomorphic as graphs and the graph isomorphism preserves the cyclic order of the edges incident to a vertex.

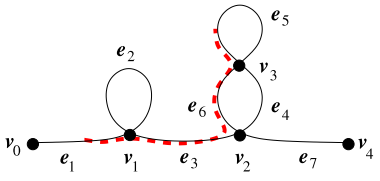

Fig. 5. Assembly graph with two endpoints v_0 and v_4 and three 4-valent rigid vertices v_1, v_2, and v_3.

Example 4.1. Consider the graph Γ given in Fig. 5. We have $E(v_2) = \{e_3, e_4, e_6, e_7\}$, and a cyclic order (e_3, e_6, e_4, e_7) can be associated to the vertex v_2. Thus $(v_2, [e_3e_6e_4e_7])$ is a 4-valent rigid vertex of Γ. The edges e_3 and e_4 are neighbors to e_6 and e_7 with respect to v_2.

A *transverse path* is a path in Γ without repeating edges where each pair of consecutive edges are non-neighbors. A transverse path "goes straight" through each 4-valent vertex it visits. If a transverse path starts and ends with two endpoints of Γ, it is called a *linear component* of Γ. If a transverse path starts and ends at a single 4-valent vertex, it is called a *circular component* of Γ. An *oriented component* is a component where the edges are oriented in the direction of the transverse path. If all components of Γ are oriented, we say that Γ is *oriented*. If $\gamma = (v_0, e_1, v_1, e_2, \ldots, e_m, v_m)$ is a path in Γ, the reverse path $(v_m, e_m, \ldots, v_1, e_1, v_0)$ is denoted with $\overline{\gamma}$.

Two linear transverse paths with endpoints are *equivalent* if they are either identical, or one is the reverse of the other. Two transverse paths γ, γ' without endpoints are *equivalent* if they have the same cyclic order: $[\gamma] = [\gamma'] = [\overline{\gamma}]$.

An assembly graph Γ is called *simple* if there is a transverse Eulerian path in Γ, meaning that there is a transverse path γ that contains every edge from Γ exactly once. In other words, an assembly graph is simple if it consists of a single transverse component. For example, the assembly graph Γ in Fig. 5 is simple.

An *open path* is a path that does not contain the end vertices. An open path in Γ can be seen as a homeomorphic image (in topological sense) of the open interval $(0, 1)$ in Γ. A *polygonal path* is an open path with non-repeating vertices whose every two consecutive edges are neighbors. In Fig. 5, a Hamiltonian polygonal path is indicated by a dotted line.

A set of pairwise disjoint paths $\{\gamma_1, \ldots, \gamma_k\}$ in Γ is called *Hamiltonian* if their union contains all 4-valent vertices of Γ. A path γ is called *Hamiltonian* if the set $\{\gamma\}$ is Hamiltonian. Hamiltonian polygonal paths are of special interest, since they model the correct order of the MDSs in the macronuclear DNA. For more details, see [4]. An assembly graph is called *realizable* if it has a Hamiltonian polygonal path.

A *smoothing* of a 4-valent vertex in an oriented graph Γ can be seen as a removal of the vertex and its neighborhood, followed by attaching two parallel arcs, as shown in Fig. 6. If the smoothing is orientation preserving then it is *parallel smoothing* (or *p-smoothing*, as in Fig. 6 left). Otherwise, the smoothing is *non-parallel smoothing* (or *n-smoothing* as in Fig. 6 right). Although we start with an oriented assembly graph, after smoothing, there are no predefined edge directions for the resulting graph.

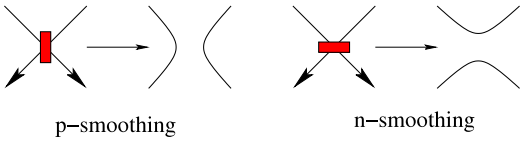

Fig. 6. Two types of smoothing, parallel (p-)smoothing (left) and non-parallel (n-)smoothing (right).

Let Γ be a simple realizable assembly graph, and let γ be a polygonal Hamiltonian path. At each vertex of Γ, a smoothing is determined from γ in such a way that γ stays "intact" after the smoothing; see Fig. 7. Such a smoothing is called *the smoothing of Γ with respect to γ*, and the resulting graph is denoted by $\tilde{\Gamma}_\gamma$ (see [4] for details). Note that $\tilde{\Gamma}_\gamma$ does not contain any 4-valent vertices, and might have two vertices of degree 1 corresponding to the endpoints of Γ. The graph $\tilde{\Gamma}_\gamma$ can have more than one component. If Γ has endpoints, then one of the components in $\tilde{\Gamma}_\gamma$ is an arc (linear) component, and the rest are closed curves (topologically circles).

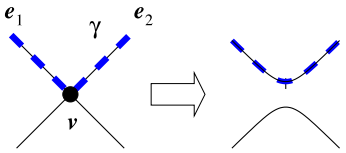

Fig. 7. Smoothing with respect to a polygonal path.

Paths in 4-regular graphs and their properties have been studied previously (for example, see [6,14]). The transverse paths that we use here are similar to Eulerian circuits obtained by γ-decomposition of the edges that appear in [14], as well as within the concepts of gene assembly in [9]. In addition, a notion similar to the smoothings with respect to Hamiltonian sets discussed in this section is found as σ-decompositions of edges also defined in [14]. However, these concepts in [14] are defined and studied for graphs without end points and for circuit decompositions, and, in our case, the presence of end points of assembly graphs and polygonal paths give essential differences in the problems discussed.

4.2. Smoothing of Γ and rewriting rules

Let Γ be an oriented assembly graph, and let $\{C_1, \ldots C_n\}$ be the set of all transverse components in Γ. Each component C_i for $i \in \{1, 2, \ldots n\}$ is uniquely determined by a transverse path γ_i that contains every edge of C_i exactly once. Let $\gamma_i = (v_0^i, e_1^i, v_1^i, e_2^i, \ldots, e_n^i, v_{n_i}^i)$. We assign a linear word $w_{\gamma_i} = e_1^i e_2^i \cdots e_n^i$ to γ_i if C_i is a linear component with two endpoints ($v_0^i \neq v_{n_i}^i$), or a circular word $[w_{\gamma_i}] = [e_1^i e_2^i \cdots e_n^i]$ if C_i is a circular component with no endpoints ($v_0^i = v_{n_i}^i$). If an edge e with endpoints v, v' is oriented from v to v' with a transverse path γ, we write $\bar{e}$ for the same edge in $w_{\overline{\gamma}}$. Suppose that Γ consists of k linear transverse components $\{C_1, \ldots, C_k\}$ and $n - k$ circular components $\{C_{k+1}, \ldots, C_n\}$.

Definition 4.2. The set of words $W_\Gamma = \{w_{\gamma_1}, \ldots, w_{\gamma_k}, [w_{\gamma_{k+1}}], \ldots, [w_{\gamma_n}]\}$ is called a *phrase* of Γ.

Remark 4.3. Let Γ be an assembly graph that contains exactly one transverse component with two endpoints and n components with no endpoints. Then the phrase of Γ contains a single linear word, and all other words are circular. The rewriting operations of insertion, deletion, and inversion defined on sets of intermediates can be also defined, and they are closed on phrases that contain exactly one linear word.

Proposition 4.4. *Let Γ be an oriented assembly graph that contains exactly one linear transverse component C (possibly with other circular transverse components). If $\tilde{\Gamma}_{\{v\}}$ is the assembly graph obtained from Γ by a smoothing of a vertex v that belongs to C, then the phrase $W_{\tilde{\Gamma}_{\{v\}}}$ is obtained from the phrase W_Γ by a single operation of deletion, insertion, or inversion.*

Proof. Let Γ be an oriented assembly graph. Let $\{C, C_1, \ldots, C_s\}$ be the set of all transverse components in Γ such that C is the only linear transverse component with two endpoints i and t. Then, $W_\Gamma = \{w_\gamma, [w_{\gamma_1}], \ldots, [w_{\gamma_s}]\}$ is the phrase of Γ.

Let v be a vertex in C. There are two possibilities.

(i) The vertex v belongs only to C (Fig. 8(A)).
(ii) There is a component C_i with no endpoints such that v belongs to both C and C_i (Fig. 8(B)).

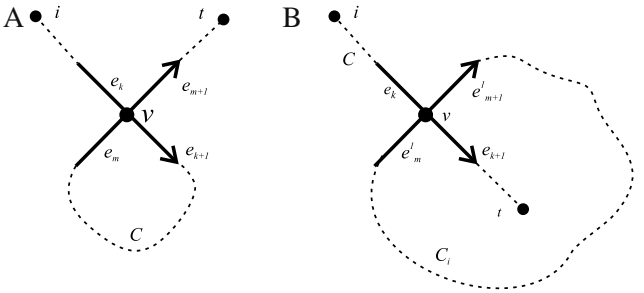

Fig. 8. (A) The rigid vertex v belongs to a single transverse component C. (B) The rigid vertex v belongs to two transverse components C and C_i.

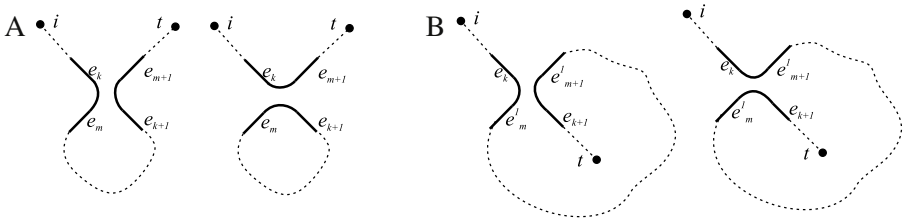

Fig. 9. Smoothings of the vertex v corresponding to (A) and (B) in Fig. 8, respectively.

In case (i), the rigid vertex $(v, [e_k e_m e_{k+1} e_{m+1}])$ is such that

$$\gamma = (i, e_1, v_1, \ldots, e_k, v, e_{k+1}, \ldots, e_m, v, e_{m+1}, \ldots, e_n, t)$$

is a transverse path that corresponds to C (Fig. 8(A)).

If the smoothing of v is non-parallel, then the transverse path γ transforms to

$$\gamma' = (i, e_1, v_1, \ldots, e_k, v, e_m, \ldots, e_{k+1}, v, e_{m+1}, \ldots, e_n, t),$$

where the portion of γ from e_m to e_{k+1} is reversed (Fig. 9(A) left). Therefore, the phrase of $\tilde{\Gamma}_{\{v\}}$,

$$W_{\tilde{\Gamma}_{\{v\}}} = \{e_1 e_2 \cdots e_k \overline{e_m \cdots e_{k+1}} e_{m+1} \cdots e_n, [w_{\gamma_1}], \ldots, [w_{\gamma_s}]\},$$

is obtained from W_Γ by an inversion. By parallel smoothing of v, the transverse component C is divided into two transverse components in $W_{\tilde{\Gamma}_{\{v\}}}$; one is linear, with a transverse path $(i, e_1, \ldots, e_k, v, e_{m+1}, \ldots, e_n, t)$, and the other is circular, with a transverse path $(v, e_{k+1}, \ldots, e_m, v)$ (Fig. 9(A) right). Therefore, the phrase of $\tilde{\Gamma}_{\{v\}}$

$$W_{\tilde{\Gamma}_{\{v\}}} = \{e_1 e_2 \cdots e_k e_{m+1} \cdots e_n, [e_{k+1} \cdots e_m], [w_{\gamma_1}], \ldots, [w_{\gamma_s}]\}.$$

is obtained from W_Γ by deletion.

In case (ii), the rigid vertex $(v, [e_k e^i_m e_{k+1} e^i_{m+1}])$ belongs to two transverse components C and, say, C_i (see Fig. 8(B)). The transverse components C and C_i are determined by the transverse paths

$$\gamma = (i, e_1, v_1, e_2, \ldots, e_k, v, e_{k+1}, \ldots, e_n, t) \quad \text{and} \quad \gamma_1 = (v, e^i_{m+1}, \ldots, e^i_m, v),$$

respectively. (Note that $[e^i_{m+1} \cdots e^i_m] = [\overline{e^i_m \cdots e^i_{m+1}}]$, since C_i is circular.)

If $\tilde{\Gamma}_{\{v\}}$ is obtained by smoothing v, then C_i is inserted into the linear component C (Fig. 9(B)).

Then the phrase of $\tilde{\Gamma}_{\{v\}}$ becomes

$$W_{\tilde{\Gamma}_{\{v\}}} = \{e_1 e_2 \cdots e_k w e_{k+1} \cdots e_n, [w_{\gamma_2}], \ldots, [w_{\gamma_s}]\},$$

where $w = \overline{e^i_m \cdots e^i_{m+1}}$ or $w = e^i_{m+1} \cdots e^i_m$ for the case of non-parallel or parallel smoothing of v, respectively. It is clear that $W_{\tilde{\Gamma}_{\{v\}}}$ is obtained from W_Γ by an insertion in both cases. □

Let $W = \{w\}$ for $w = I_0N_1I_1N_2 \cdots I_{k-1}N_kI_k$ be a minimal set of intermediates over the alphabet $\mathfrak{A}_k$, where $N_i \in \mathcal{M}$ and $I_j \in \mathcal{I}$. We briefly recall the construction of an assembly graph $\Gamma = \Gamma_W$ as described in [3]. The set of vertices in Γ_W is $\{1, 2 \ldots, k, k+1\}$. Edges are uniquely labeled by symbols in $\mathfrak{A}_k$ that appear in W, but we identify the edges with these symbols. All edges are labeled by single symbols, except for edges that have labels M_1 or M_k, which may be labeled by two or three symbols. More details follow in the next section (and see Example 4.6 and Fig. 11 below). In a manner similar to Proposition 4.4, we have the converse.

Proposition 4.5. *Let W be a minimal set of intermediates, and let Γ_W be the corresponding assembly graph. If W' is a set of intermediates obtained from W by a single operation of deletion, insertion, or inversion that increases the degree of w, then the graph $\Gamma_{W'}$ is obtained from Γ_W by a parallel or non-parallel smoothing of a vertex.*

Comparing to Proposition 4.4, we require that an operation is degree increasing in Proposition 4.5. If an operation is not degree increasing, the graph may not contain a corresponding vertex to smooth it.

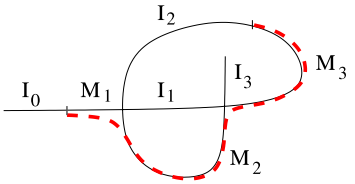

Fig. 10. The operation 'deletion of I_2' has no corresponding smoothing of a vertex.

For example, in Fig. 10, an assembly graph corresponding to the minimal set of intermediates $W = \{I_0M_1I_1M_3I_2M_2I_3\}$ is depicted. If we apply the deletion of I_2 (which does not increase the degree), we obtain $W' = \{I_0M_1I_1M_3M_2I_3, [I_2]\}$. However, there in no smoothing of the graph that gives rise to the word W'.

4.3. Smoothing strategies and sets of intermediates

Let Γ be a simple realizable assembly graph, let γ be a polygonal Hamiltonian path, and let S be a subset of vertices in Γ. The *S-partial smoothing of Γ with respect to γ* is an assembly graph with a set of 4-valent vertices $V(\Gamma) \setminus S$, denoted by $\tilde{\Gamma}_{(\gamma,S)}$, obtained by smoothing of all vertices in S with respect to γ. The S-partial smoothing Γ with respect to γ is *successful* if the transverse path γ (the corresponding path after smoothings) stays in the linear component.

Let $\mathbb{S} = \{S_1, \ldots, S_h\}$ be an ordered partition of $V(G)$. We say that $\mathbb{S}$ is a *smoothing strategy* for Γ with respect to γ, where γ is a Hamiltonian polygonal path in Γ. We call $\mathbb{S}$ *successful* if S_i is successful in $\tilde{\Gamma}_{(\gamma,S_i')}$, for every $i = 1, \ldots, h$, where $S_1' = \emptyset$ and $S_i' = \cup_{j=1}^{i-1} S_j$.

Given a minimal set of intermediates $W = \{w\}$ for $w = I_0N_1I_1N_2 \cdots I_{k-1}N_kI_k$, we describe the construction of an assembly graph $\Gamma = \Gamma_W$. We first look at the appearances of M_1 and M_k in W. If $M_kI_bM_1 \sqsubset W$ or $\overline{M_1}I_b\overline{M_k} \sqsubset W$, then $M_kI_bM_1$ is the label of an edge in Γ with endpoints $\{2, k\}$. If the appearances of M_1 and M_k are not separated by a single symbol in w, we have two special edges, as follows. If M_1 (respectively, M_k) is not reversed in W, we set an edge I_bM_1 (respectively, $\underline{M_kI_t}$) with endpoints $\{s+1, 2\}$ if $M_sI_bM_1 \sqsubset W$ or $\{s, 2\}$ if $\overline{M_s}I_bM_1 \sqsubset W$ (respectively, $\{k, s\}$ if $M_kI_tM_s \sqsubset W$ or $\{k, s+1\}$ if $M_kI_t\overline{M_s} \sqsubset W$). Otherwise, if M_1 (respectively, M_k) is reversed in W, we set an edge $\overline{M_1}I_b$ (respectively, $\underline{I_t\overline{M_k}}$) with endpoints $\{2, s\}$ if $\overline{M_1}I_bM_s \sqsubset W$ or $\{2, s+1\}$ if $\overline{M_1}I_b\overline{M_s} \sqsubset W$ (respectively, $\{s+1, k\}$ if $M_sI_t\overline{M_k} \sqsubset W$ or $\{s, k\}$ if $\overline{M_s}I_t\overline{M_k} \sqsubset W$). For all other vertices $j \neq b, t$, the symbols I_j and $N_i \notin \{M_1, \overline{M_1}, M_k, \overline{M_k}\}$ are edges of Γ_W. The endpoints of N_s are $\{s, s+1\}$, where $N_s = M_s$ or $N_s = \overline{M_s}$. The endpoints of I_j ($j \neq 0, b, t$) are defined as follows:

$$\begin{aligned}
&\{s+1, s'\} && \text{if} \quad M_sI_jM_{s'} \sqsubset W,\\
&\{s, s'\} && \text{if} \quad \overline{M_s}I_jM_{s'} \sqsubset W,\\
&\{s+1, s'+1\} && \text{if} \quad M_sI_j\overline{M_{s'}} \sqsubset W,\\
&\{s, s'+1\} && \text{if} \quad \overline{M_s}I_j\overline{M_{s'}} \sqsubset W.
\end{aligned}$$

The endpoints of I_0 are $\{1, s\}$ if $I_0M_s \sqsubset W$, or $\{1, s+1\}$ if $I_0\overline{M_s} \sqsubset W$, and similarly for I_k. At every vertex, the orientations of the edges specifying the rigidity of the vertex are obtained such that the edges representing consecutive symbols in w are non-neighbors.

Example 4.6. Consider the actin I gene of *O. trifallax* with the corresponding minimal set of intermediates $W = \{I_0M_3I_1M_4I_2M_6I_3M_5I_4M_7I_5M_9I_6M_{10}I_7\overline{M_2}I_8M_1I_9M_8I_{10}\}$. The vertex set of Γ_W is $\{1, 2 \ldots, 10, 11\}$. The set of edges is constructed as follows. The symbols M_1 and M_{10} are not reversed in W, and therefore there are two special edges in Γ_W, I_8M_1 and $M_{10}I_7$ (see Fig. 11). The endpoints of I_8M_1 (respectively, $M_{10}I_7$) are $\{2, 2\}$ (respectively, $\{10, 3\}$), since $\overline{M_2}I_8M_1 \sqsubset W$ (respectively, $M_{10}I_7\overline{M_2} \sqsubset W$). All other symbols in W represent distinct edges in Γ_W. The edges labeled by N_s (where $N_s = M_s$ or $N_s = \overline{M_s}$) have endpoints $\{s, s+1\}$ for every $s \in \{2, 3, \ldots k-1\}$, as shown in Fig. 11. The edges I_0 and I_{10} have endpoints $\{1, 3\}$ and

$\{9, 11\}$, respectively, because $I_0M_3 \sqsubset W$ and $M_8I_{10} \sqsubset W$. Finally, for all remaining j ($j \in \{1, 2, 3, 4, 5, 6, 9\}$), the edge I_j has endpoints $\{s+1, s'\}$, where $M_sI_jM_{s'} \sqsubset W$. For instance, I_2 is the edge $\{5, 6\}$, since $M_4I_2M_6 \sqsubset W$. The rigidity of each vertex v is specified by ordering the incident edges so that consecutive symbols in W are non-neighboring edges in Γ_W with respect to v.

Remark 4.7. The construction described in this section gives a simple assembly graph Γ_W with a single linear component which consists of a transverse path w. The assembly graph from Example 4.6 in Fig. 11 is a simple assembly graph with transverse component $I_0M_3I_1M_4I_2M_6I_3M_5I_4M_7I_5M_9I_6M_{10}I_7\overline{M_2}I_8M_1I_9M_8I_{10} = w$.

In [4], it is proved that $\Gamma = \Gamma_W$ contains a Hamiltonian polygonal path γ_Γ with a subpath $M_1M_2 \cdots M_k$. The assembly graph Γ_W in Example 4.6 contains a Hamiltonian polygonal path $\gamma_\Gamma = (I_8M_1M_2 \cdots M_9M_{10}I_7)$.

Let $S \subset V(\Gamma)$ be a set of vertices. Suppose that the assembly graph obtained by partial smoothing Γ_S contains transverse components $\{C, C_1, \ldots, C_s\}$ such that C is the only linear transverse component, and all others are circular. Let $\{\gamma, \gamma_1, \ldots, \gamma_s\}$ be the corresponding transverse paths, and let $W_{\Gamma_S} = \{w_\gamma, [w_{\gamma_1}], \ldots, [w_{\gamma_s}]\}$ be the phrase of Γ_S. Then the phrase $W_{\Gamma_S} = \{w_\gamma, [w_{\gamma_1}], \ldots, [w_{\gamma_s}]\}$ is a set of intermediates. We tie the smoothing strategies with the assembly strategies in the following theorem.

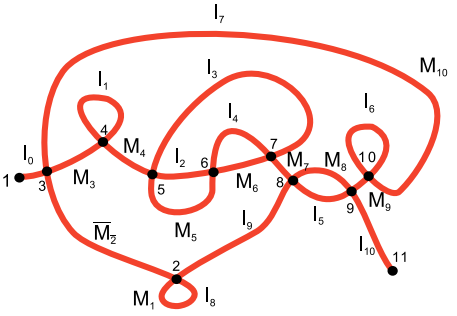

Fig. 11. An assembly graph $\Gamma = \Gamma_W$ that corresponds to the minimal set of intermediates $W = \{I_0M_3I_1M_4I_2M_6I_3M_5I_4M_7I_5M_9I_6M_{10}I_7\overline{M_2}I_8M_1I_9M_8I_{10}\}$.

Proposition 4.8. *Let $W = \{w\}$ be a realizable minimal set of intermediates over the alphabet $\mathfrak{A}_k$, let $\Gamma = \Gamma_W$ be the corresponding simple assembly graph, and let $\gamma = \gamma_\Gamma$ be a Hamiltonian polygonal path with subpath $M_1 \cdots M_k$. If a nested sequence $\emptyset = S_0 \subset S_1 \subset S_2 \subset \cdots \subset S_h = V(\Gamma)$ is a successful smoothing strategy for Γ with respect to γ, then the sequence of phrases $(W_\Gamma, W_{\tilde{\Gamma}_{(\gamma,S_1)}}, W_{\tilde{\Gamma}_{(\gamma,S_2)}}, \ldots, W_{\tilde{\Gamma}_{(\gamma,S_h)}})$ is an assembly strategy.*

Proof. Inductively, suppose that $W_j = W_{\Gamma_{(\gamma,S_j)}}$ is a set of intermediates and that $\mathrm{dg}(W_{j-1}) < \mathrm{dg}(W_j)$ for all $j < i$. Let $S'_i = S_i \setminus S_{i-1}$ for $i = 1, \ldots, h$. Then $W_{\tilde{\Gamma}_{(\gamma,S_i)}}$ is obtained from $W_{\tilde{\Gamma}_{(\gamma,S_{i-1})}}$ by smoothing vertices in S'_i with respect to γ. Let $(c_1, \ldots, c_d)$ be an ordered sequence of elements (vertices) of S'_i. Let $W'_0 = W_{\tilde{\Gamma}_{(\gamma,S_{i-1})}}$, $W'_1 = W_{\tilde{\Gamma}_{(\gamma,S_{i-1}\cup\{c_1\})}}$, and inductively, $W'_j = W_{\tilde{\Gamma}_{(\gamma,S_{i-1}\cup\{c_1,\ldots,c_j\})}}$, and $W'_d = W_{\tilde{\Gamma}_{(\gamma,S_{i-1}\cup\{c_1,\ldots,c_d\})}} = W_{\tilde{\Gamma}_{(\gamma,S_i)}}$.

Then, for each j, W'_j is obtained from W'_{j-1} by smoothing c_j with respect to γ. By Proposition 4.4, the set of intermediates W'_j is obtained from W'_{j-1} by a single operation of insertion, deletion, or inversion. There are two neighboring edges incident to c_j labeled by M_s and M_{s-1} for some $s \in \{1, 2, \ldots, k\}$. Since c_j is smoothed with respect to the Hamiltonian polygonal path γ, the smoothing results in an edge with label $M_{s-1}M_s$, so that $\mathrm{dg}(W'_j) > \mathrm{dg}(W'_{j-1})$. Therefore, $W_{\tilde{\Gamma}_{(\gamma,S_i)}}$ is obtained from $W_{\tilde{\Gamma}_{(\gamma,S_{i-1})}}$ by a sequence of insertion, deletion, or inversion operations, each of which increases the degree. Hence, $W_{\tilde{\Gamma}_{(\gamma,S_{i-1})}} \Rightarrow^* W_{\tilde{\Gamma}_{(\gamma,S_i)}}$ for every $i \in \{1, 2, \ldots, k\}$.

The set of intermediates W_Γ is minimal by construction. The set of intermediates $W_{\tilde{\Gamma}_{(\gamma,S_h)}}$ corresponds to a completely smoothed graph $\tilde{\Gamma}_{(\gamma,S_h)}$, and thus either $M_1M_2 \cdots M_k \sqsubset W_{\tilde{\Gamma}_{(\gamma,S_h)}}$ or $\overline{M_kM_{k-1} \cdots M_1} \sqsubset W_{\tilde{\Gamma}_{(\gamma,S_h)}}$. In any case, the degree of $W_{\tilde{\Gamma}_{(\gamma,S_h)}}$ is $k-1$, i.e., $W_{\tilde{\Gamma}_{(\gamma,S_h)}}$ is maximal. Hence, the sequence $(W_\Gamma, W_{\tilde{\Gamma}_{(\gamma,S_1)}}, \ldots, W_{\tilde{\Gamma}_{(\gamma,S_h)}})$ is an assembly strategy. □

We observe that the converse of Proposition 4.8 does not hold. There are assembly strategies that do not correspond to successful smoothing strategies. In some assembly strategies, there are sets of intermediates where MDSs appear in distinct words, which is not allowed in successful smoothing strategies (or any intramolecular model). One such example is the assembly strategy following the experimental findings of rearrangements of actin I gene of *O. trifallax* in the following section. This fact shows that the assembly strategies on words defined here are more general than the successful smoothing strategies defined on graphs. In other words, using assembly strategies on words, one can easily simulate any gene assembly strategy assuming intramolecular rearrangement.

On the other hand, if one considers an arbitrary smoothing strategy (not necessarily successful), then we obtain the following.

Proposition 4.9. *The sequence $W_0 \Rightarrow^* W_1 \Rightarrow^* \cdots \Rightarrow^* W_n$ is an assembly strategy for the set of intermediates W if and only if the partition of the set of vertices $\{V(\Gamma_{W_i}) \setminus V(\Gamma_{W_{i+1}}) : i = 0, \ldots, n-1\}$ such that W_i is the phrase of Γ_{W_i} defines a smoothing strategy for Γ_{W_0}.*

5. Application of sets of intermediates to experimental data

In [17], the assembly of the *O. trifallax* actin I gene was discussed in terms of the partially assembled molecules experimentally detected. We apply our method of assembly strategies to the putative intermediate molecules based on these experimental data to show that, theoretically, there are two distinct largest pathways for descrambling the *O. trifallax* actin I gene. The micronuclear actin I gene with approximately proportional sequence length is schematically represented in Fig. 1. Recall that dark segments represent MDS sequences as labeled, and white segments represent IESs. It is also assumed that there is an IES segment I_0 to the left of M_3, and another IES segment I_{10} to the right of M_8. The segment M_2 is inverted.

The set of partially assembled molecules observed in [17] is listed below. Due to the PCR extraction process, these molecules are subsequences of larger molecules. The notation a through f follows the notation in [17]. The analysis follows.

$$\mathsf{a} = \overline{M_1 I_8} M_2 M_3 I_1 M_4 I_2 M_6 I_3 M_5 I_4 M_7 I_5 M_9,$$
$$\mathsf{b} = \overline{M_1 I_8} M_2 M_3 I_1 M_4 I_2 M_6 M_7 I_5 M_9,$$
$$\mathsf{c} = \overline{M_1 I_8} M_2 M_3 M_4 M_5 M_6 M_7 M_8 M_9,$$
$$\mathsf{d} = \overline{M_8 I_9 M_1} I_8 M_2 M_3 M_4 M_5 I_4 M_7 I_5 M_9,$$
$$\mathsf{e} = \overline{M_8 I_9 M_1} I_8 M_2 M_3 M_4 I_2 M_6 M_7 I_5 M_9,$$
$$\mathsf{f} = \overline{M_8 I_9 M_1} I_8 M_2 M_3 I_1 M_4 I_2 M_6 I_3 M_5 I_4 M_7 I_5 M_9.$$

Based on this data, we construct sets of intermediates that model the micronuclear and macronuclear gene and the possible intermediate molecules. Let

$$W = \{I_0 M_3 I_1 M_4 I_2 M_6 I_3 M_5 I_4 M_7 I_5 M_9 I_6 M_{10} I_7 \overline{M_2} I_8 M_1 I_9 M_8 I_{10}\},$$
$$A = \{I_0 \overline{I_7 M_{10} I_6}\ \bar{\mathsf{a}}\ I_9 M_8 I_{10}\} = \{I_0 \overline{I_7 M_{10} I_6}\ \bar{\mathsf{f}}\ I_{10},\},$$
$$B = \{I_0 \overline{I_7 M_{10} I_6}\ \bar{\mathsf{b}}\ I_9 M_8 I_{10}, [I_3 M_5 I_4]\},$$
$$C = \{I_0 \overline{I_7 M_{10} I_6}\ \bar{\mathsf{c}}\ I_9 I_5 I_{10}, [I_1], [I_3 I_2 I_4]\},$$
$$D = \{I_0 \overline{I_7 M_{10} I_6}\ \bar{\mathsf{d}}\ I_{10}, [I_1], [I_2 M_6 I_3]\},$$
$$E = \{I_0 \overline{I_7 M_{10} I_6}\ \bar{\mathsf{e}}\ I_{10}, [I_1], [I_3 M_5 I_4]\},$$
$$W' = \{I_0 \overline{I_7 M_{10} M_9 M_8 M_7 M_6 M_5 M_4 M_3 M_2 M_1 I_8} I_9 I_5 I_{10}, [I_1], [I_6], [I_3 I_2 I_4]\}$$

be sets of intermediates over alphabet $\mathfrak{A}_{10}$. Then, W models the scrambled micronuclear actin I gene and W' models the assembled macronuclear actin I gene. Note that W is minimal and W' is maximal with respect to the partial order $\Rightarrow^*$.

We have $a, f \sqsubset A$, and, similarly, the sets of intermediates B, C, D, E model the partially assembled molecules b, c, d, e, respectively, in the following sense: these sets of intermediates A, B, C, D, E contain (reverses of) a, (f), b, c, d, e as subwords, and satisfy the following lemma. Our assumption is that a and f are part of the same set of intermediates because of their significant overlap.

Lemma 5.1. *The sets of intermediates A through F satisfy the following properties.*

(1) $W \Rightarrow^* v$ *for any* $v \in \{A, B, C, D, E\}$.
(2) $v \Rightarrow^* W'$ *for any* $v \in \{A, B, C, D, E\}$.

Proof. (1) We have $W_0 \overset{inv}{\Rightarrow} A$, since A is obtained from W_0 by inverting the portion $M_3 \cdots I_7$, and $\mathrm{dg}(W_0) = 0 < \mathrm{dg}(A) = 1$. Similarly,

B is obtained from A by deleting $I_3 M_5 I_4$,
D is obtained from A by deleting I_1 and $I_2 M_6 I_3$,
E is obtained from B by deleting I_1,

and each deletion increases the degree by 1. Let

$$W_1 = \{I_0 \overline{I_7 M_{10} I_6 M_9 I_5 M_7 M_6 M_5 I_4 I_3 I_2 M_4 M_3 M_2 I_8} M_1 I_9 M_8 I_{10}, [I_1]\},$$
$$W_2 = \{I_0 \overline{I_7 M_{10} I_6 M_9 I_5 M_7 M_6 M_5 M_4 M_3 M_2 I_8} M_1 I_9 M_8 I_{10}, [I_1], [I_4 I_3 I_2]\},$$
$$W_3 = \{I_0 I_7 M_{10} I_6 M_9 M_8 I_9 \overline{M_1} I_8 M_2 M_3 M_4 M_5 M_6 M_7 I_5 I_{10}, [I_1], [I_4 I_3 I_2]\}.$$

Then W_1 is obtained from E by insertion of $[\underline{M_5 I_4} I_3]$, W_2 is obtained from W_1 by a deletion of $[I_4 I_3 I_2]$, and W_3, C are obtained from W_2, W_3, by inverting $\overline{I_5 M_7} \cdots I_9 M_8$ and $\overline{I_9 M_1} \cdots M_7$, respectively. Furthermore, we have

$$\mathrm{dg}(E) = 3 < \mathrm{dg}(W_1) = 4 < \mathrm{dg}(W_2) = 5 < \mathrm{dg}(W_3) = 6 < \mathrm{dg}(C) = 7;$$

hence $E \Rightarrow^* C$. Similarly, we have $D \Rightarrow^* C$ through insertion of $[M_6 I_2 I_3]$, deletion of $I_2 I_3 I_4$, and two inversions of $M_8 I_9 \cdots \overline{I_5}$ and $I_9 M_1 \cdots M_7$.

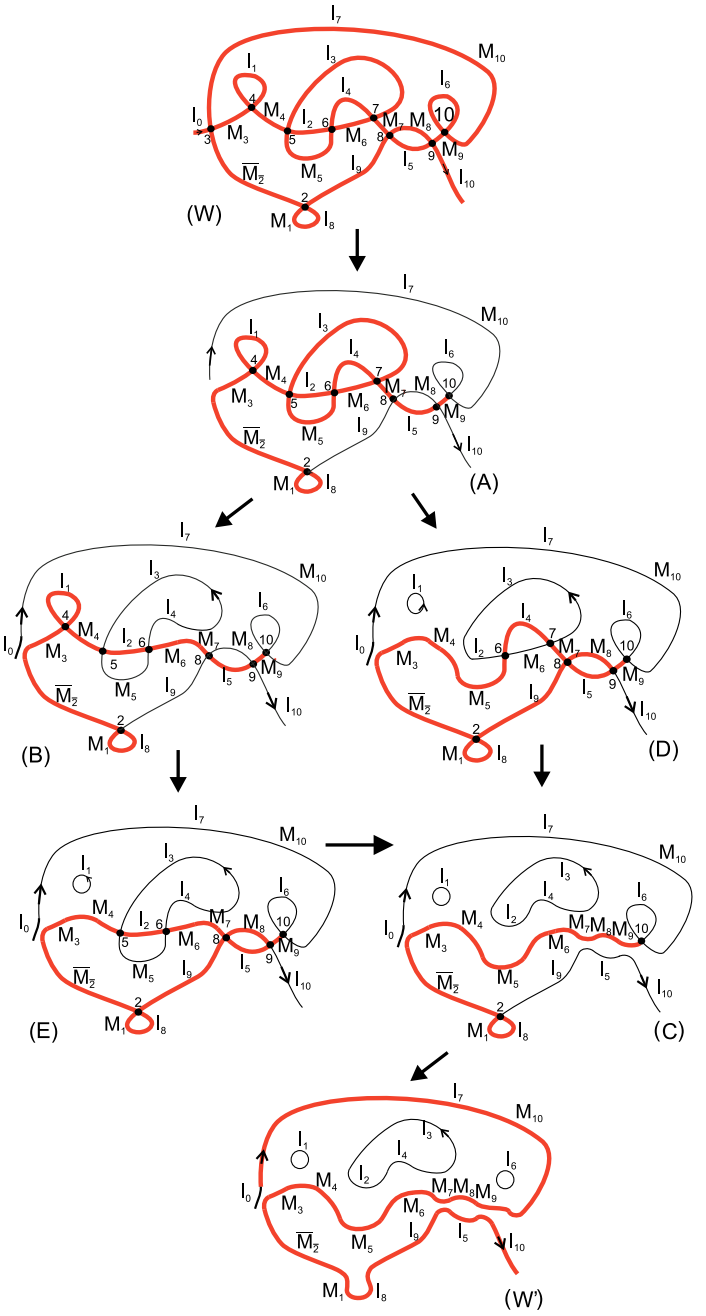

Fig. 12. Two smoothing strategies (W, A, B, E, C, W') and (W, A, D, C, W') as models of two different pathways for actin I macronuclear gene assembly. The putative intermediate molecules extracted experimentally are indicated with bold red. Both proposed strategies have a set of intermediates with MDSs separated in distinct molecules (B, E, and D). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(2) We only need to prove that $C \Rightarrow^* W'$. Let

$$V = \{I_0\overline{I_7M_{10}I_6M_9M_8M_7M_6M_5M_4M_3M_2M_1I_8}I_9I_5I_{10}, [I_1], [I_3I_2I_4]\}.$$

The set of intermediates V is obtained from C by inversion of I_8M_1, and W' is obtained from V by deletion of I_6. In addition, $\mathrm{dg}(C) = 7 < \mathrm{dg}(V_1) = 8 < \mathrm{dg}(W') = 9$. Therefore, $C \Rightarrow^* W'$. □

Proposition 5.2. *There are exactly two largest subsets of* $\{W, A, B, C, D, E, W'\}$ *that form assembly strategies.*

Proof. The degrees of the sets of intermediates W, W' and A through E are

$$\mathrm{dg(W)} = 0, \quad \mathrm{dg(A)} = 1, \quad \mathrm{dg(B)} = 2, \quad \mathrm{dg(C)} = 7, \quad \mathrm{dg(D)} = 3, \quad \mathrm{dg(E)} = 3, \quad \mathrm{dg(W')} = 9.$$

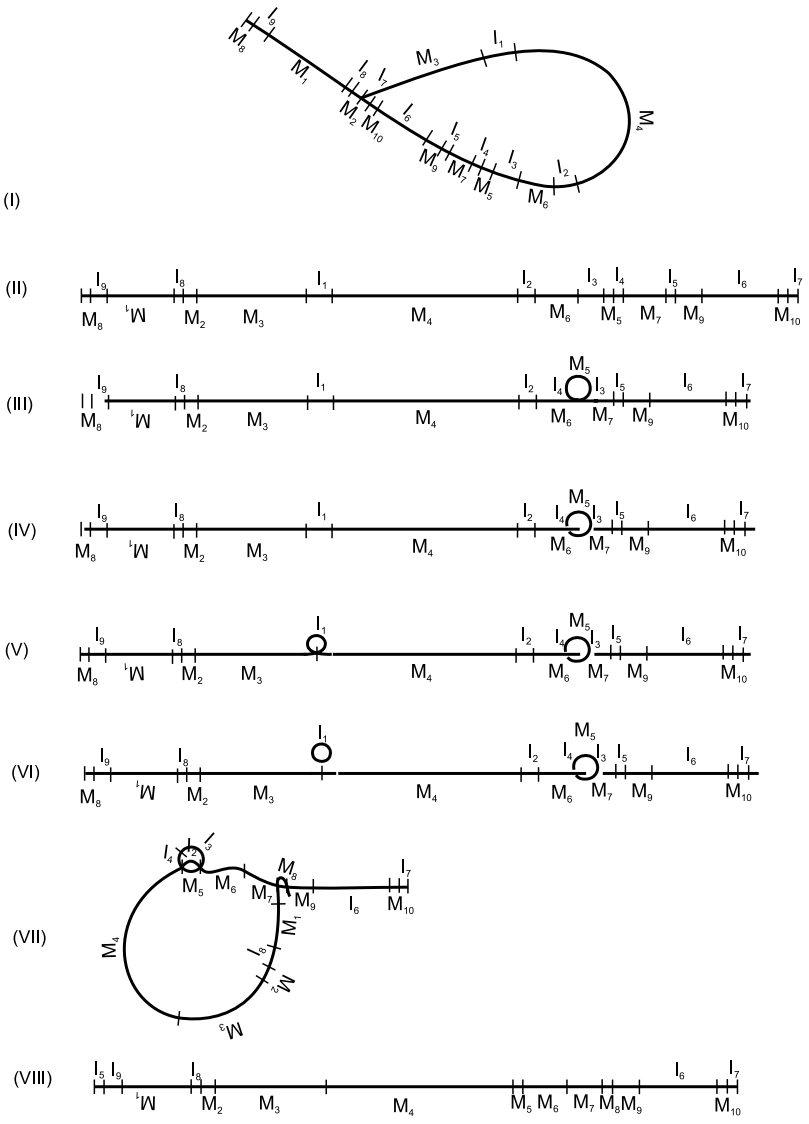

Fig. 13. First proposed pathway for unscrambling the actin I gene in *O. trifallax*, based on the possible intermediates observed in [17].

According to the degrees of the sets of intermediates, there are two sequences that are possible candidates for assembly strategies. They are

$$W \Rightarrow^* A \Rightarrow^* D \Rightarrow^* C \Rightarrow^* W',$$
$$W \Rightarrow^* A \Rightarrow^* B \Rightarrow^* E \Rightarrow^* C \Rightarrow^* W'.$$

By Lemma 5.1, we have $W \Rightarrow^* A$ and $C \Rightarrow^* W'$. We also checked in Lemma 5.1 that $A \Rightarrow^* B$, $B \Rightarrow^* E$, $E \Rightarrow^* C$, and $A \Rightarrow^* D$, $D \Rightarrow^* C$. On the other hand, $B \not\Rightarrow D$ and $D \not\Rightarrow B$. The latter case is obvious, since $\mathrm{dg(D)} = 3 > \mathrm{dg(B)} = 2$. To show that $B \not\Rightarrow D$, assume the contrary. If $B \Rightarrow D$, then there is a sequence of sets of intermediates $B = W_0, W_1, \ldots, W_h = D$ such that, for each i ($i = 1, \ldots, h$), W_i is obtained from W_{i-1} by a single operation of deletion, insertion, or inversion, and $\mathrm{dg}(W_{i-1}) \leq \mathrm{dg}(W_i)$. Since the assembled segment M_6M_7 satisfies $M_6M_7 \sqsubset B$ and $M_6M_7 \not\sqsubset D$, there is some $j \in \{2, 3, \ldots, h\}$ such that $M_6M_7 \sqsubset W_{j-1}$ and $M_6M_7 \not\sqsubset W_j$. The set of intermediates W_j is obtained from W_{j-1} by a single deletion, inversion, or insertion operation, so, by the structure of B and D, that is only possible if $\mathrm{dg}(W_{j-1}) > \mathrm{dg}(W_j)$, a contradiction. Therefore, $B \not\Rightarrow D$. □

Note that any proper substrategy of $W \Rightarrow^* A \Rightarrow^* D \Rightarrow^* C \Rightarrow^* W'$ or $W \Rightarrow^* A \Rightarrow^* B \Rightarrow^* E \Rightarrow^* C \Rightarrow^* W'$ can be viewed as an assembly strategy on its own (but not the largest). In that case, one can form a total of 19 assembly strategies.

In Fig. 12, pathways corresponding to the two assembly strategies are described in terms of assembly graphs and smoothings. The assembly graph representation of the actin I MIC gene is given in Fig. 12 (W). Note that the path γ : $M_1 - M_2 - \cdots - M_9 - M_{10}$ is a Hamiltonian polygonal path. The graphs (W, A, B, E, C, W') determine one pathway, and

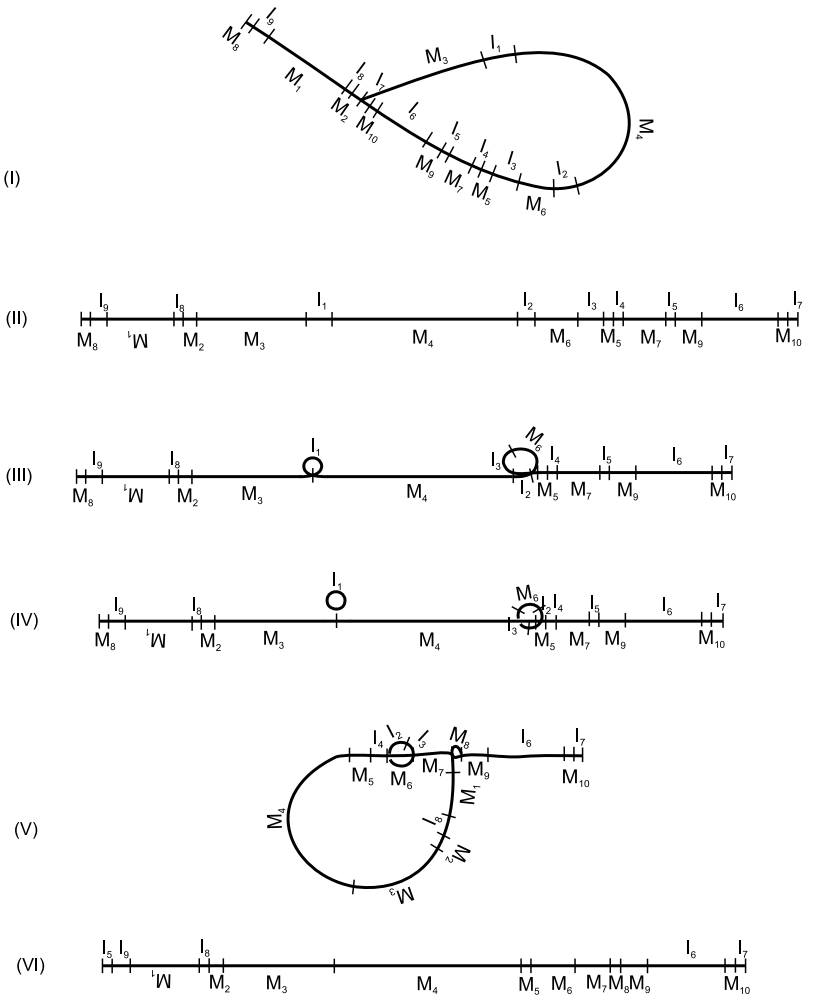

Fig. 14. Second proposed pathway for unscrambling the actin I gene in *O. trifallax*, based on the possible intermediates observed in [17].

(W, A, D, C, W') another. The first case corresponds to the sequence of smoothings at crossings {3}, {7}, {4}, {5, 6, 8, 9}, {2, 10}, and the latter to {3}, {4, 5}, {6, 7, 8, 9}, {2, 10}.

In [3], it is proposed that, if a DNA molecule forms a structure corresponding to an assembly graph, then the recombination can happen simultaneously, corresponding to smoothing all vertices of the assembly graph. Considering the experimental data containing partially recombined molecules with either MDS_5 or MDS_6 but not both, we postulate that the assembly graph structure represented as (W) in Fig. 12 does not necessarily form. To obtain a more accurate physical model reflecting the two pathways, we depict the recombination processes by partially assembled spatial graphs in Figs. 13 and 14. Figures are to scale in the sense that the lengths of line segments are proportional to the numbers of nucleotides of DNA segments.

6. Conclusion

The combinatorial model introduced here uses sets of formal words (called sets of intermediates) to describe all molecules present at a given step of the gene rearrangement. The DNA recombination is modeled by three rewriting rules: insertion, deletion, and inversion. The collection of all sets of intermediates together with the three rewriting rules forms a rewriting system, which we show is confluent. The gene assembly process can be viewed as a successive application of rewriting rules to an initial singleton set of intermediates that models the scrambled MIC gene. Using such an approach, one can generate all possible pathways for rearrangement. Based on the partially processed molecules that are experimentally observed, the model predicts only a restricted number of different pathways. For instance, we exhibit two assembly strategies for actin I in *O. trifallax*. Subsequent biological experiments motivated by the present model performed on TEBPα gene of *O. trifallax* showed new possible intermediate (circular) molecules including IES-IES junctions (see [2]).

Mathematical models for DNA rearrangement have been introduced previously (e.g., [4,11,15]). The rewriting rules applied on sets of intermediates are built upon the operations defined in [15]. The authors of [11] assume that all MDSs

must be present on a single molecule at each step of the rearrangement, while our model views each intermediate step as a collection of multiple different molecules. The experimental results in [2,17] support the later approach.

Namely, detected sequences of partially rearranged molecules reported in [2] also confirm that the rearrangement cannot be completely intramolecular, as modeled earlier (e.g., [4,9–11]). In addition, in [2], circular molecules were experimentally observed, supporting the use of multiple circular words included in the set of intermediates.

We show that our model generalizes the model on assembly graphs defined in [4]. In addition, the sets of intermediates can easily be ordered to form rearrangement pathways, and can be incorporated into algorithmic computations.

Although our model provides a concise symbolic way to describe the rearrangement of molecules, it has disadvantages; the model does not offer explanations on mechanisms whereby these molecules are kept together. Considerations on spatial structures to address this issue in conjunction with experimental verifications are desirable.

Even though we use certain genera of ciliates (*Oxytricha* and *Stylonychia*) as model organisms to study gene recombination and gene assembly, the theoretical model proposed here can be applied to any site-specific DNA rearrangement processes observed in nature.

Acknowledgments

We are grateful to B.P. Higgins and L.F. Landweber for helpful conversations. N.J. and M.S. were supported in part by NSF Grant DMS #0900671. We are grateful to the referees for valuable comments.

References

[1] A. Angeleska, Combinatorial models for DNA rearrangement, Ph.D. Thesis, Univ. of South Florida, Tampa FL, 2009.
[2] A. Angeleska, B.P. Higgins, N. Jonoska, L.F. Landweber, Predicting DNA Rearrangement Pathways (in preparation).
[3] A. Angeleska, N. Jonoska, M. Saito, L.F. Landweber, RNA-guided DNA assembly, Journal of Theoretical Biology 248 (4) (2007) 706–720.
[4] A. Angeleska, N. Jonoska, M. Saito, DNA recombinations through assembly graphs, Discrete Applied Mathematics 157 (14) (2009) 3020–3037.
[5] G.T. Bignell, T. Santarius, J.C. Pole, A.P. Butler, J. Perry, E. Pleasance, C. Greenman, A. Menzies, S. Taylor, S. Edkins, P. Campbell, M. Quail, B. Plumb, L. Matthews, K. McLay, P.A. Edwards, J. Rogers, R. Wooster, P.A. Futreal, M.R. Stratton, Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution, Genome Res. 17 (:9) (2007) 1296–1303.
[6] A. Bouchet, *k*-transformations, complementations, switching, in: G. Hahn, et al. (Eds.), Cycles and Rays, in: NATO ASI Series C, Vol. 301, 1987, pp. 41–50. 1296–1303.
[7] P.J. Campbell, P.J. Stephens, E.D. Pleasance, S. O'Meara, H. Li, T. Santarius, L.A. Stebbings, C. Leroy, S. Edkins, C. Hardy, J.W. Teague, A. Menzies, I. Goodhead, D.J. Turner, C.M. Clee, M.A. Quail, A. Cox, C. Brown, R. Durbin, M.E. Hurles, P.A.W. Edwards, G.R. Bignell, M.R. Stratton, P.A. Futreal, Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-ensequencing, Nature Genetics 40 (2008) 722–729.
[8] W-J. Chang, S. Kuo, L.F. Landweber, A new scrambled gene in the ciliate *Uroleptus*, Gene 368 (2006) 72–77.
[9] A. Ehrenfeucht, T. Hajru, G. Rozenberg, Gene Assembly through cyclic graph decomposition, Theoretical Computer Science 281 (2002) 325–349.
[10] A. Ehrenfeucht, I. Petre, DM. Prescott, G. Rozenberg, Cilcularity and other invariants of gene assembly in ciliates, in: M. Ito, G. Paun, S. Yu (Eds.), Words, Semigroups, and Transductions, World Scientific, 2001, pp. 81–97.
[11] A. Ehrenfeucht, T. Harju, I. Petre, D.M. Prescott, G. Rozenberg, Computing in Living Cells, Springer, 2005.
[12] W. Harriman, H. Volk, N. Defranoux, M. Wabl, Immunoglobulin class switch recombination, Annu Rev Immunol. 11 (1993) 361384.
[13] C.H. Bassing, W. Swat, F.W. Alt, The mechanism and regulation of chromosomal V(D)J recombination, Cell 109 (2002) 4555.
[14] A. Kotzig, Eulerian lines in finite 4-valent graphs and their transformations, in: Theory of Graphs, Academic Press, New York, 1968, pp. 219–230.
[15] L.F. Landweber, L. Kari, The evolution of cellular computing: natures solution to a computational problem, BioSystems 52 (1999) 3–13.
[16] L.F. Landweber, L. Kari, Universal molecular computation in ciliates, in: L.F. Landweber, E. Winfree (Eds.), Evolution as Computation, Springer, Berlin Heidelberg New York, 2002, pp. 257–274.
[17] M. Mollenbeck, Y. Zhou, A.R.O. Cavalcanti, F. Jonsson, W.-J. Chang, S. Juranek, T.G. Doak, G. Rozenberg, H.J. Lipps, L.F. Landweber, The pathway for detangling a scrambled gene, PLoS ONE 3 (6) (2008) :e2330.
[18] M. Nowacki, V. Vijayan, Y. Zhou, T. Doak, E. Swart, L.F. Landweber, RNA-template guided DNA recombination: epigenetic reprogramming of a genome rearrangement pathway, Nature 451 (2008) 153–158.
[19] I. Petre, Invariants of gene assembly in stichotrichous ciliates, IT Oldenbourg Wissenschftsverlag 48 (2006) 161–167.
[20] D. Prescott, The evolutionary scrambling and developmental unscrambling of germline genes in hypotrichous ciliates, Nucleic Acid Research 27 (5) (1999).
[21] D.M. Prescott, The DNA of ciliated protozoa, Microbiological Reviews 58 (2) (1994) 233–267.
[22] D.M. Prescott, A.F. Greslin, Scrambled actin I gene in the micronucleus of Oxytricha nova, Dev Genet. 13 (1) (1992) 66–74.
[23] D.M. Prescott, A. Ehrenfeucht, G. Rozenberg, Template-guided recombination for IES elimination and unscrambling of genes in stichotrichous ciliates, J. of Theoretical Biology 222 (2003) 323–330.
[24] S.L. Tausta, L.A. Klobutcher, Detection of circular forms of eliminated DNA during macronuclear development in *E. crassus*, Cell 59 (6) (1989) 1019–1026.